Homework 01

STAT 430, Fall 2017

Due: Friday, September 15, 11:59 PM

Please see the homework instructions document for detailed instructions and some grading notes. Failure to follow instructions will result in point reductions.

Exercise 1

[10 points] This question will use data in a file called `hw01-data.csv`. The data contains four predictors: `a`, `b`, `c`, `d`, and a response `y`.

After reading in the data as `hw01_data`, use the following code to test-train split the data.

```
set.seed(42)
train_index = sample(1:nrow(hw01_data), size = round(0.5 * nrow(hw01_data)))
train_data = hw01_data[train_index, ]
test_data = hw01_data[-train_index, ]
```

Next, fit four linear models using the training data:

- Model 1: `y ~ .`
- Model 2: `y ~ . + I(a ^ 2) + I(b ^ 2) + I(c ^ 2)`
- Model 3: `y ~ . ^ 2 + I(a ^ 2) + I(b ^ 2) + I(c ^ 2)`
- Model 4: `y ~ a * b * c * d + I(a ^ 2) + I(b ^ 2) + I(c ^ 2)`

For each of the models above, report:

- Train RMSE
- Test RMSE
- Number of Parameters, Excluding the Variance

To receive full marks, arrange this information in a well formatted table. Also note which model is best for making predictions.

[Not Graded] For fun, find a model that outperforms each of the models above. *Hint:* Consider some exploratory data analysis. *Hint:* Your instructor's solution uses a model with only seven parameters. Yours may have more.

Exercise 2

[10 points] For this question we will use the `Boston` data from the `MASS` package. Use `?Boston` to learn more about the data.

```
library(readr)
library(tibble)
library(MASS)
data(Boston)
Boston = as_tibble(Boston)
```

Use the following code to test-train split the data.

```
set.seed(42)
boston_index = sample(1:nrow(Boston), size = 400)
train_boston = Boston[boston_index, ]
test_boston  = Boston[-boston_index, ]
```

Fit the following linear model that uses `medv` as the response.

```
fit = lm(medv ~ . ^ 2, data = train_boston)
```

Fit two additional models, both that perform worse than `fit`, with respect to prediction. One should be a smaller model, relative to `fit`. The other should be a larger model, relateive to `fit`. Call them `fit_smaller` and `fit_larger` respectively. Any "smaller" model should be nested in any "larger" model.

Report these three models as well as their train RMSE, test RMSE, and number of parameters. Note: you may report the models used using their `R` syntax. To receive full marks, arrange this information in a well formatted table.

Exercise 3

[10 points] How do outliers affect prediction? Usually when fitting regression models for explanation, dealing with outliers is a complicated issue. When considering prediction, we can empirically determine what to do.

Continue using the `Boston` data, training split, and models from Exercise 2. Consider the model stored in `fit` from Exercise 2. Obtain the standardized residuals from this fitted model. Refit this model with each of the following modifications:

- Removing observations from the training data with absolute standardized residuals greater than 2.
- Removing observations from the training data with absolute standardized residuals greater than 3.

(a) Use these three fitted models, including the original model fit to unmodified data, to obtain test RMSE. Summarize these results in a table. Include the number of observations removed for each. Which performs the best? Were you justified modifying the training data?

(b) Using the *best* of these three fitted models, create a 99% **prediction interval** for a new observation with the following values for the predictors:

crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	black	lstat
0.02763	75.0	3.95	0	0.4280	6.595	22.8	5.4011	3	252	19.3	395.63	4.32